Ehsani and Drabløs *BMC Bioinformatics* (2016) 17:296
DOI 10.1186/s12859-016-1160-0

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

TopoICSim: a new semantic similarity measure based on gene ontology

Rezvan Ehsani[1,2] and Finn Drabløs[1*]

Abstract

Background: The Gene Ontology (GO) is a dynamic, controlled vocabulary that describes the cellular function of genes and proteins according to tree major categories: biological process, molecular function and cellular component. It has become widely used in many bioinformatics applications for annotating genes and measuring their semantic similarity, rather than their sequence similarity. Generally speaking, semantic similarity measures involve the GO tree topology, information content of GO terms, or a combination of both.

Results: Here we present a new semantic similarity measure called TopoICSim (Topological Information Content Similarity) which uses information on the specific paths between GO terms based on the topology of the GO tree, and the distribution of information content along these paths. The TopoICSim algorithm was evaluated on two human benchmark datasets based on KEGG pathways and Pfam domains grouped as clans, using GO terms from either the biological process or molecular function. The performance of the TopoICSim measure compared favorably to five existing methods. Furthermore, the TopoICSim similarity was also tested on gene/protein sets defined by correlated gene expression, using three human datasets, and showed improved performance compared to two previously published similarity measures. Finally we used an online benchmarking resource which evaluates any similarity measure against a set of 11 similarity measures in three tests, using gene/protein sets based on sequence similarity, Pfam domains, and enzyme classifications. The results for TopoICSim showed improved performance relative to most of the measures included in the benchmarking, and in particular a very robust performance throughout the different tests.

Conclusions: The TopoICSim similarity measure provides a competitive method with robust performance for quantification of semantic similarity between genes and proteins based on GO annotations. An R script for TopoICSim is available at http://bigr.medisin.ntnu.no/tools/TopoICSim.R.

Keywords: Gene ontology, Semantic similarity measure, Tree topology

Background

Gene ontology

The Gene Ontology (GO) is a useful resource in bioinformatics that provides structured and controlled vocabularies to describe protein function and localization according to three general categories: biological process (BP), molecular function (MF), and cellular component (CC) [1, 2]. Each of these three annotation categories is structured as its own rooted Directed Acyclic Graph (rDAG). An rDAG is a treelike data structure with a unique root node, the relationships between nodes are directed (oriented), and the structure is non-recursive, i.e. without cycles.

The GO consortium updates on a regular basis a GO Annotation (GOA) [3] database with new GO terms that are linked to genes and gene products by relevant studies. GO is widely used in several bioinformatics applications, including gene functional analysis of DNA microarray data [4], gene clustering [5], disease similarity [6], and prediction and validation of protein-protein interactions [7].

Each GO annotation is assigned together with an evidence code (EC) that refers to the process used to assign the specific GO term to a given gene [8]. All ECs are

* Correspondence: finn.drablos@ntnu.no
[1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, P.O. Box 8905, NO-7491 Trondheim, Norway
Full list of author information is available at the end of the article

reviewed by a curator, except ECs assigned with the Inferred from Electronic Annotation (IEA) code.

Semantic similarity

Measuring similarity between objects that share some attributes is a central issue in many research areas such as psychology, information retrieval, biomedicine, and artificial intelligence [9, 10]. Such similarity measures can be based on comparing features that describe the objects, and a semantic similarity measure uses the relationships which exist between the features of the items being compared [11]. Blanchard et al. have established a general model for comparing semantic similarity measures based on a subsumption hierarchy [12]. They divide tree-based similarities into two categories: those based only on the hierarchical relationships between the terms [13], and those combining additional statistics such as term frequency in a corpus [14].

In a biological perspective, the functional similarity term was proposed to describe the similarity of genes or gene products as the similarity between their GO annotation terms. To establish a suitable functional similarity between genes has become an important aspect of many biological studies. For example have previous studies shown that there is a correlation between gene expression and GO semantic similarity [15].

Since GO terms are organized as an rDAG, the functional similarity can be estimated by a semantic similarity. Pesquita et al. have proposed a general definition of semantic similarity between genes or gene products [16]. Here a semantic similarity is a function which, given two sets of terms annotating two biological entities, returns a numerical value presenting the closeness in meaning between them. This similarity measure is based on comparing all possible pairs of the two sets of GO terms, or selective subsets of them.

Comparing terms

In general measuring the similarity between two terms can be divided into three main categories: edge-based, node-based and hybrid methods. The edge-based approaches are based on counting the number of edges in the specific path between two terms. In most edge-based measures, a distance function is defined on the shortest path (*SP*) or on the average of all paths [17, 18]. This distance can easily be converted into a similarity measure. Such approaches rely on two assumptions which are seldom true in biological reality. First that nodes and edges are uniformly distributed, and second that edges at the same level in the GO graph correspond to identical distances between terms. Node-based measures are based on the information content (*IC*) of the terms involved. The *IC* value gives a measure of how specific and informative a term is. The *IC* is relying on the probability of terms occurring in a corpus, and Resnik [19] used the negative logarithm of the likelihood of a term to quantify its *IC*.

$$IC(t) = -logp(t) \quad (1)$$

This definition leads to higher *IC* for terms with lower frequency. Obviously, *IC* values increase as a function of depth in the GO graph (this is illustrated in the presentation of TopoICSim, in Results). Resnik used the maximal value among all common ancestors between two terms as a similarity measure, i.e., the *IC* of the lowest common ancestor (LCA) [19]. Since the similarity value of Resnik's measure is not limited to one (1.0), Lin [14] and Jiang [20] proposed their methods to normalize the similarity value between 0.0 and 1.0. Most node-based methods are based on Resnik's measure which only considers the *IC* of a single common ancestor and ignores the information on paths in subgraphs composed from common ancestors and pairs GO terms. So, hybrid methods have been proposed to account for both nodes and edges in the subgraph. For example Wang et al. introduced a similarity measure combining the structure of the GO graph with the *IC* values, integrating the contribution of all terms in a GO subgraph, including all the ancestors [21].

Comparing genes or gene products

Genes are normally annotated using several terms within a particular GO category (MF, BP or CC). Thus, with an available measure function to compute similarity of terms, it is necessary to define an aggregated similarity measure to compare sets of terms. Generally these measures can be divided into two categories: pairwise and groupwise methods [16].

Pairwise approaches measure similarity between two genes by combining the similarities between their terms. Some approaches apply all possible pairwise combination of terms from the two sets, whereas others consider only the best-matching pair for each term. The final similarity between two genes is then defined by combining these pairwise similarities, mostly by the average, the maximum, or the sum [3, 19].

Groupwise methods are not based on combing similarities between individual terms, but rather compute gene similarities by one of three main approaches: set, graph, or vector. In set approaches the similarity is computed by set techniques on the annotations. Graph-based similarity measures calculate similarity between genes using graph matching techniques where each gene is presented as subgraphs of GO terms. And finally, in vector approaches each gene is represented in vector space with each term corresponding to a dimension. Similarity can

be estimated using vector-based similarity measures, mostly cosine similarity [22].

Existing measures

For presentation of some existing methods we introduce the following definitions. Suppose g_1 and g_2 are two given genes or gene products annotated by two sets of GO terms $\{t_{11}, t_{12}, ..., t_{1n}\}$ and $\{t_{21}, t_{12}, ..., t_{2m}\}$. The first measure we will introduce is IntelliGO [22], which is a vector-based method. Each gene is represented as a vector $g = \sum_i \alpha_i e_i$ where $\alpha_i = w(g, t_i)IFA(t_i)$, and where $w(g, t_i)$ represents the weight assigned to the evidence code between g and t_i, and $IFA(t_i)$ is the invers annotation frequency of the term t_i. Here e_i is the i-th basis vector corresponding to the annotation term t_i. The dot product between two gene vectors is defined as in (2) and (3).

$$g_1 * g_2 = \sum\nolimits_{i,j} \alpha_i \beta_j e_i * e_j \tag{2}$$

$$e_i * e_j = \frac{2Depth(LCA)}{MinSPL(t_{1i}, t_{2j}) + 2Depth(LCA)} \tag{3}$$

Here $Depth(LCA)$ is the depth of the deepest common ancestor for t_{1i}, t_{2j} and $MinSPL(t_{1i}, t_{2j})$ is the length of the shortest path between t_{1i}, t_{2j} which passes through LCA. The similarity measure for two genes vectors g_1 and g_2 is then defined using the cosine formula (4).

$$SIM_{IntelliGO}(g_1, g_2) = \frac{g_1 * g_2}{\sqrt{g_1 * g_1}\sqrt{g_2 * g_2}} \tag{4}$$

The second measure presented here was introduced by Wang et al. [21]. They considered for the different contributions that terms are related by *is_a* and *part_of*. The semantic contribution that ancestor terms make to a child term is estimated by (5).

$$SV(t) = \sum_{x \in Anc(t)} S_t(x) \tag{5}$$

Here $S_t(t) = 1$ and $S_t(x) = \max\{w_e * S_t(t_i) | t_i \in children\ of(x)\}$, where $w_e \in [0, 1]$ is a value that corresponds to the semantic contribution factor for edge e, and $childrenof(x)$ returns the immediate children of x that are ancestors of t and $S_t(t_i) = \prod_{x \in P(t, t_{i-1})} max\ w_k$ where $P(t, t_{i-1})$ is the path between t and t_{i-1}. They used the weights $w_{is\_a} = 0.8$ and $w_{part\_of} = 0.6$. Then they defined the similarity of two terms as in (6).

$$S(t_{1i}, t_{2j}) = \frac{\sum_{x \in ComAnc(t_{1i}, t_{2j})} S_{t_{1i}}(x) + S_{t_{2j}}(x)}{SV(t_{1i}) + SV(t_{2j})} \tag{6}$$

Finally the Wang measure uses a best-matched approach (BMA) to calculate similarity between two genes according to (7).

$$SIM_{Wang}(g_1, g_2) = \frac{\sum_{i=1}^{n} max_j S(t_{1i}, t_{2j}) + \sum_{j=1}^{m} max_i S(t_{1i}, t_{2j})}{n + m} \tag{7}$$

The third measure is Lord's measure [3], which is based on Resnik's similarity. The Resnik similarity is defined as in (8).

$$SIM_{Resnik}(t_{1i}, t_{2j}) = IC(LCA(t_{1i}, t_{2j})) \tag{8}$$

The Lord measure is estimated as the average of the Resnik similarity over all t_{1i} and t_{2j}.

$$SIM_{Lord}(g_1, g_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} SIM_{Resnik}(t_{1i}, t_{2j})}{n \times m} \tag{9}$$

The next measure was introduced by Al-Mubaid et al. [23]. First they calculate the length of all shortest paths (PLs) for all (t_{1i}, t_{2j}) pairs. Then the average on the PLs defines the distance between two genes g_1 and g_2 as in (10).

$$PL(g_1, g_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} PL(t_{1i}, t_{2j})}{n \times m} \tag{10}$$

Finally they use function (11) to convert the distance to a similarity value.

$$SIM_{Mubaid}(g_1, g_2) = e^{-0.2 \times PL(g_1, g_2)} \tag{11}$$

The last measure presented here is SimGIC [24], which also is called the Weighted Jaccard measure. Let G_1 and G_2 be the GO terms and their ancestors for two genes g_1 and g_2, respectively. The SimGIC is defined as the ratio between the sum of the ICs of terms in the intersection and the sum of the ICs of terms in the union (12).

$$SimGIC(g_1, g_2) = \frac{\sum_{t \in G_1 \cap G_2} IC(t)}{\sum_{t \in G_1 \cup G_2} IC(t)} \tag{12}$$

We will now describe the implementation and testing of a new method, TopoICSim, and compare it to the measures introduced above using several different test data sets. In this measure we have tried to decrease any bias induced by irregularity of the rDAG. In particular, TopoICSim examines all common ancestors for a pair of GO terms, and not only the last (or deepest) common ancestor, which is the case for the measures introduced above. Details regarding the evaluation measures, the datasets and approaches that were used for benchmarking and the actual implementation are given in Methods.

Methods

IntraSet similarity and discriminating power

To evaluate TopoICSim relative to existing methods we first used two different benchmarks based on the GO

properties studied by Benabderrahmane et al. [22]. For the KEGG benchmark they used a diverse set of 13 human KEGG pathways. The assumption when testing the KEGG dataset is that genes belonging to a specific pathway share a similar biological process, so the estimated similarity was based on BP annotations (Table 1). They also defined a Pfam benchmark, using data from the Sanger Pfam database [25] for 10 different Pfam human clans. The assumption when testing Pfam clans is that genes belonging to a specific clan share a similar molecular function, so the estimated similarity was based on MF annotations (Table 1).

They used two measures, *IntraSet Similarity* and *Discriminating Power* on the benchmark datasets to evaluate their method. Let S be a collection of genes where $S = \{S_1, S_2, \ldots, S_p\}$ (each S_k can be e.g. a Pfam clan or a KEGG pathway). For each S_k, let $\{g_{k1}, g_{k2}, \ldots, g_{kn}\}$ be the set of n genes in S_k. *IntraSet* similarity is a measure to calculate the average similarity over all pairwise comparisons within a set of genes (13).

$$IntraSetSim(S_k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} Sim\left(g_{ki}, g_{kj}\right)}{n^2} \quad (13)$$

InterSet similarity can be estimated for two sets of genes S_k and S_l composed of n and m genes, respectively, as the average of all similarities between pairs of genes from each of the two sets S_k and S_l (14).

$$InterSetSim(S_k, S_l) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} Sim\left(g_{ki}, g_{lj}\right)}{n \times m} \quad (14)$$

The ratio of the *IntraSet* and *InterSet* average gene similarities can be defined as the discriminating power (*DP*) (15).

$$DP_{Sim}(S_k) = \frac{(p-1)IntraSetSim(S_k)}{\sum_{i=1, i\neq k}^{p} InterSetSim(S_k, S_i)} \quad (15)$$

It is important to have high *IntraSet* similarity and at the same time high *Discriminating Power* for a measure. Therefore we decided to define a new measure, *IntraSet Discriminating Power* (*IDP*), using the following formula (16).

$$IDP_{Sim}(S_k) = IntraSetSim(S_k) \times DP_{Sim}(S_k) \quad (16)$$

The *IDP* value estimates the ability to identify similarity between gene sets in a dataset, and at the same time discriminate these sets from other genes in the dataset.

We compared the results obtained with our TopoICSim method with the five existing state-of-the-art similarity measures described in the introduction. For the benchmark datasets, *IntraSet, DP,* and *IDP* values were calculated by our method and compared to those estimated using the other measures.

Expression similarity

Many recent studies have shown that genes that are biologically and functionally related often maintain this similarity both in their expression profiles as well as in their GO annotations [15]. To test this assumption we selected three sets of genes from the Hallmark datasets, which is a collection of 50 gene sets representing specific well-defined biological processes [26]. These three gene

Table 1 List of human KEGG pathways and Pfam clans used for benchmarking

	KEGG			Pfam	
Pathway	Name	#genes	Accession	Name	#genes
hsa00040	Pentose and glucuronate interconversions	26	CL0099.10	ALDH-like	18
hsa00920	Sulfur metabolism	13	CL0106.10	6PGD_C	8
hsa00140	C21-Steroid homone metabolism	17	CL0417.1	BIR-like	9
hsa00290	Valine, leucine and isoleucine biosynthesis	11	CL0165.8	Cache	5
hsa00563	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	23	CL0149.9	CoA-acyltrans	7
hsa00670	One carbon pool by folate	16	CL0085.11	FAD_DHS	12
hsa00232	Caffeine metabolism	7	CL0076.9	FAD_Lum_binding	18
hsa03022	Basal transcription factors	38	CL0289.3	FBD	6
hsa03020	RNA polymerase	29	CL0119.10	Flavokinase	7
hsa04130	SNARE interactions in vesicular transport	38	CL0042.9	Flavoprotein	10
hsa03450	Non-homologous end-joining	14			
hsa03430	Mismatch repair	23			
hsa04950	Maturity onset diabetes of the young	25			
Total #genes		280			100

These datasets were obtained directly from [22]

sets are labeled as G2M_CHECKPOINT, DNA_REPAIR, and IL6_JAK_STAT3_SIGNALING, with 200, 151, and 87 genes respectively. The expression values for the genes across multiple cell types and experiments have been obtained from FANTOM5 [27] using the "CAGE peak based expression table (RLE normalized) of robust CAGE peaks for human samples with annotation" file. The expression values were listed according to clusters of transcriptional start sites, therefore some genes were initially assigned multiple expression values, corresponding to unique clusters of start sites. We combined expression values for each gene and then transformed the total expression by log2. Each gene could then be represented as a vector with 1829 expression values.

We used three expression similarities (Pearson correlation, Spearman correlation, and Distance correlation (*DC*)), against the three annotation similarities (TopoICSim, IntelliGO, and Wang) that showed the best performance during initial testing (see Results).

Previous studies have shown that in most cases there is no meaningful correlation when pairs of individual genes are used to estimate correlation between expression and annotation similarities, but that this can be improved by grouping methods, comparing groups or clusters of genes [15]. In these methods, the gene pairs are split into groups of equal intervals according to the annotation (or expression) similarity values between the gene pairs. Then correlation between expression and annotation similarities is defined as correlation between the average of these similarities on the splits [28, 29]. There are many reasons for poor correlation when interactions between individual genes are considered. For example, genes may be involved in multiple and different processes across a dataset. Comparison of individual genes will underestimate similarity due to these differences, whereas grouping methods can highlight shared properties within groups. We therefore decided to group results by using a Self-Organizing Map (SOM) algorithm on (r, s) pairs, where r and s are one of the expression and annotation similarities respectively. A SOM is a topology-preserving mapping of high-dimensional data based on artificial neural networks. It consists of a geometry of nodes mapped into a k-dimensional space, initially at random, which is iteratively adjusted. In each iteration the nodes move in the direction of selected data points, where the movement depends upon the distances to the data points, so that data points located close to a given node have a larger influence than data points located far away. Thereby, neighboring points in the initial topology tend to be mapped to close or identical nodes in the k-dimensional space [30]. We calculated correlation between expression and annotation similarities for all clusters and then identified clusters showing good correlation. Final correlation is reported as average correlation of individual expression and annotation similarities within these clusters. This approach was applied to all possible combination of (r, s) values, i.e., 9 combinations in total.

Distance correlation

Distance Correlation (*DC*) as introduced by Székely and Bakirov [31] is a method to estimate the dependency between two random variables. It measures the discrepancy between the joint function and the product of its marginal functions in a specific weighting scheme in L_2 space. More strictly, let (X, Y) be a pair of random variables with joint function $f_{(X, Y)}$ and marginal functions f_X and f_Y. The distance covariance can be defined as the root of the following Eq. (17).

$$dcov^2(X,\ Y) = \int \left| f_{(X,\ Y)}(t,\ s) - f_X(t) f_Y(s) \right|^2 w(t,\ s)\, dt\, ds \tag{17}$$

This is on R^{p+q} where p and q are the dimension of X and Y respectively and $w(t, s)$ is the weight function. Now, the *DC* can be defined by distance covariance as in (18).

$$dcor(X,\ Y) = \frac{dcov(X, Y)}{\sqrt{dcov(X,\ X)}\sqrt{dcov(Y,\ Y)}} \tag{18}$$

It has been shown that the empirical *DC* for an iid sample $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ can be estimated as in (19–22).

$$DC(X, Y) = S_1 + S_2 - 2S_3 \tag{19}$$

$$S_1 = \frac{1}{n^2} \sum_{k,l=1}^{n} |x_k - x_l|_p |y_k - y_l|_q \tag{20}$$

$$S_2 = \frac{1}{n^2} \sum_{k,l=1}^{n} |x_k - x_l|_p \frac{1}{n^2} \sum_{k,l=1}^{n} |y_k - y_l|_q \tag{21}$$

$$S_3 = \frac{1}{n^3} \sum_{k=1}^{n} \sum_{l,m=1}^{n} |x_k - x_l|_p |y_k - y_m|_q \tag{22}$$

Some previous studies have applied *DC* on the expression level of gene sets [32, 33].

Evaluation by CESSM

Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) is an online tool [34] that enables the comparison of a given measure against 11 previously published measures based on their correlation with sequence, Pfam, and Enzyme Classification (ECC) similarities [35]. It uses a dataset of 13,430 protein pairs involving 1,039 unique proteins from various species. Protein pairs (from multiple species), GO (dated August 2010), and UniProt GO annotations (dated August 2008)

were downloaded from CESSM. The similarities for the 13,430 proteins pairs were calculated with TopoICSim and returned to CESSM for evaluation.

Implementation

The R programming language (version 3.2.2) was used for developing and running all programs. We used all the EC codes as annotation terms. The *ppiPre* (version 1.9), *GeneSemSim* (version 1.28.2), and *csbl.go* (version 1.4.1) packages were used to calculate IntelliGO, Wang, and SimGIC measures [36–38]. The *DC* values were estimated using the *energy* (version 1.6.2) package [39]. The SOM algorithm was performed with the *SOMbrero* (version 1.1) package [40]. All these packages are available within R Bioconductor [41].

Results

The TopoICSim measure

Here we introduce a new similarity measure which accounts for the distribution of IC on both shortest path between two terms and longest path from their common ancestor to root. A weighting scheme in terms of length of the paths is used to provide a more informative similarity measure. In the current version we do not use any weight scheme on the ECs codes. We use definitions of relevant concepts as follows.

A GO tree can be described as a triplet $\Lambda = (G, \Sigma, R)$, where G is the set of GO terms, Σ is the set of hierarchical relations between GO terms (mostly defined as *is_a* or *part_of*) [22], and R is a triplet (t_i, t_j, ξ), where $t_i, t_j \in G$ and $\xi \in R$ and $t_i \xi t_j$. The ξ relationship is an oriented child–parent relation. Top level node of the GO rDAG is the Root, which is a direct parent of the MF, BP, and CC nodes. These nodes are called aspect-specific roots and we refer to them as *root* in following.

A path P of length n between two terms t_i, t_j can be defined as in (23).

$$P : G \times G \rightarrow G \times G \cdots \times G = G^{n+1}; \quad P(t_i, t_j) = (t_i, t_{i+1}, ..., t_j) \tag{23}$$

Here $\forall\ s,\ i \le s < j,\ \exists\ \xi_s \in \Sigma,\ \exists\ \tau_s \in R,\ \tau_s = (t_s, t_{s+1}, \xi_s)$. Because G is an rDAG, there might be multiple paths between two terms, so we represent all paths between two terms t_i, t_j according to (24).

$$\mathcal{A}(t_i, t_j) = \underset{P}{\cup} P(t_i, t_j) \tag{24}$$

We use Inverse Information Content (*IIC*) values to define shortest and longest paths for two given terms t_i, t_j as shown in (25–27).

$$SP(t_i, t_j) = \underset{P \in A(t_i, t_j)}{\operatorname{argmin}} IIC(P) \tag{25}$$

$$LP(t_i, t_j) = \underset{P \in A(t_i, t_j)}{\operatorname{argmax}} IIC(P) \tag{26}$$

$$IIC(P) = \sum_{t \in P} \frac{1}{IC(t)} \tag{27}$$

We used a standard definition to calculate $IC(t)$ as shown in (28)

$$IC(t) = -\log \frac{G_t}{G_{Tot}} \tag{28}$$

Here G_t is the number of genes annotated by the term t and G_{Tot} is the total number of genes. The distribution of *IC* is not uniform in the rDAG, so it is possible to have two paths with different lengths but with same *IICs*. To overcome this problem we weight paths by their lengths, so the definitions in (25) and (26) can be updated according to (29) and (30).

$$wSP(t_i, t_j) = SP(t_i, t_j) \times len(P) \tag{29}$$

$$wLP(t_i, t_j) = LP(t_i, t_j) \times len(P) \tag{30}$$

Now let $ComAnc(t_i, t_j)$ be the set of all common ancestors for two given terms t_i, t_j. First we define the disjunctive common ancestors as a subset of $ComAnc(t_i, t_j)$ as in (31).

$$DisComAnc(t_i, t_j) = \{x \in ComAnc(t_i, t_j) \mid P(x, root) \cap C(x) = \varnothing\} \tag{31}$$

Here $P(x, root)$ is the path between x and *root* and $C(x)$ is set of all immediate children for x.

For each disjunctive common ancestor x in $DisComAnc(t_i, t_j)$, we define the distance between t_i, t_j as the ratio of the weighted shortest path between t_i, t_j which passes from x to the weighted longest path between x and *root*, as in (32–33).

$$D(t_i, t_j, x) = \frac{wSP(t_i, t_j, x)}{wLP(x, root)} \tag{32}$$

$$wSP(t_i, t_j, x) = wSP(t_i, x) + wSP(t_j, x) \tag{33}$$

Now the distance for two terms t_i, t_j can be defined according to (34).

$$D(t_i, t_j) = \min_{x \in DisComAnc(t_i, t_j)} D(t_i, t_j, x) \tag{34}$$

We convert distance values by the $\frac{Arctan(.)}{\pi/2}$ function, and the measure for two GO terms t_i and t_j can be defined as in (35).

$$S(t_i, t_j) = 1 - \frac{Arctan(D(t_i, t_j))}{\pi/2} \quad (35)$$

Note that *root* refers to one of three first levels in the rDAG. So if $DisComAnc(t_i, t_j) = \{root\}$ then $D(t_i, t_j) = \infty$ and $S(t_i, t_j) = 0$. Also if $t_i = t_j$ then $D(t_i, t_j) = 0$ and $S(t_i, t_j) = 1$.

Finally let $S = [s_{ij}]_{n \times m}$ be a similarity matrix for two given genes or gene products g_1, g_2 with GO terms $\{t_{11}, t_{12}, ..., t_{1n}\}$ and $\{t_{21}, t_{12}, ..., t_{2m}\}$, where s_{ij} is the similarity between the GO terms t_{1i} and t_{2j}. We use a *rcmax* method to calculate similarity between g_1, g_2, as defined in (36).

$$TopoICSim(g_1, g_2) = rcmax(S) = max\left(\frac{\sum_{i=1}^{n} \max_j s_{ij}}{n}, \frac{\sum_{j=1}^{m} \max_i s_{ij}}{m}\right) \quad (36)$$

We also tested other methods on the similarity matrix, in particular average and BMA, but in general *rcmax* gave the best performance for TopoICSim (data not shown).

The TopoICSim algorithm

The TopoICSim algorithm was implemented to estimate the similarity between two genes, taking their gene ID (currently *Entrez ID*) as input, together with parameters: a GO annotation type (MF, BP, and CC), a species, and an EC specification (default is NULL, which means using all ECs). The output is the similarity between the two genes. Pseudocode for the TopoICSim algorithm is presented in Fig. 1.

The *IC*s used to weigh the GO terms were calculated using the GOSim package (version 1.8.0) [42]. For each disjunctive the shortest path between the two GO terms was calculated by the *Dijkstra* algorithm in the RBGL package (version 1.46.0) [43] according to (25). Also the longest path between the disjunctive and *root* was calculated by the *topology sorting* algorithm [44] according to (26).

A simple example

To exemplify how TopoICSim computes the similarity between two given GO terms, we will illustrate the similarity between the two GO terms GO:0044260 and GO:0006139 as shown in Fig. 2, using the BP ontology of GO. According to (32), these GO terms have two disjunctive ancestors: GO:0071704 and GO:0044237. For GO:0071704 there are unique paths from GO:0071704 to root and from GO:0044260 and GO:0006139 to GO:0071704 (L1 and P1 in Fig. 2 respectively). Therefore, according to (32) the distance between these GO terms will be:

$$D(\text{GO:0044260}, \text{GO:0006139}, \text{GO:0071704}) = \frac{\left(\frac{1}{2.158} + \frac{1}{2.086} + \frac{1}{1.255} + \frac{1}{1.479} + \frac{1}{1.617}\right) \times 4}{\left(\frac{1}{1.255} + \frac{1}{1.098}\right) \times 2} = 2.75$$

For GO:0044237 there are two paths from GO:0044237 to root (L21 and L22) and two paths from GO:0044260 and GO:0006139 to GO:0044237 (P21 and P22). According to (25) and (26) and the IC values in Fig. 2 L22 and P22 are longest path and shortest path respectively, so distance for this case will be:

$$D(\text{GO:0044260}, \text{GO:0006139}, \text{GO:0044237}) = \frac{\left(\frac{1}{2.158} + \frac{1}{1.999} + \frac{1}{1.329} + \frac{1}{1.617}\right) \times 3}{\left(\frac{1}{1.329} + \frac{1}{0.407}\right) \times 2} = 1.076$$

Obviously the second value is the minimum, so the similarity between GO:0044260 and GO:0006139 according to (35) will be:

Input: Two gene IDs g_1, g_2
Output: TopoICSim similarity value between g_1, g_2

```
weighting rDAG by IIC values
extract the corresponded GO terms for g1, g2, namely {t11, t12, ..., t1n} and {t21, t22, ..., t2m}
S=[ ]n×m    # Similarity matrix
for i in range (1, n):
    for j in range (1, m):
        DCA←DisComAns(t1i, t2j)
        D=[ ]
        for x in DCA:
            D←D(t1i, t2j, x)      # According to (34)
        end for
        S←sij=1-Arctan(min D)/(π/2)
    end for
end for
TopoICSim(g1, g2)←rcmax(S)        # According to (36)
```

Fig. 1 Pseudocode for the TopoICSim algorithm

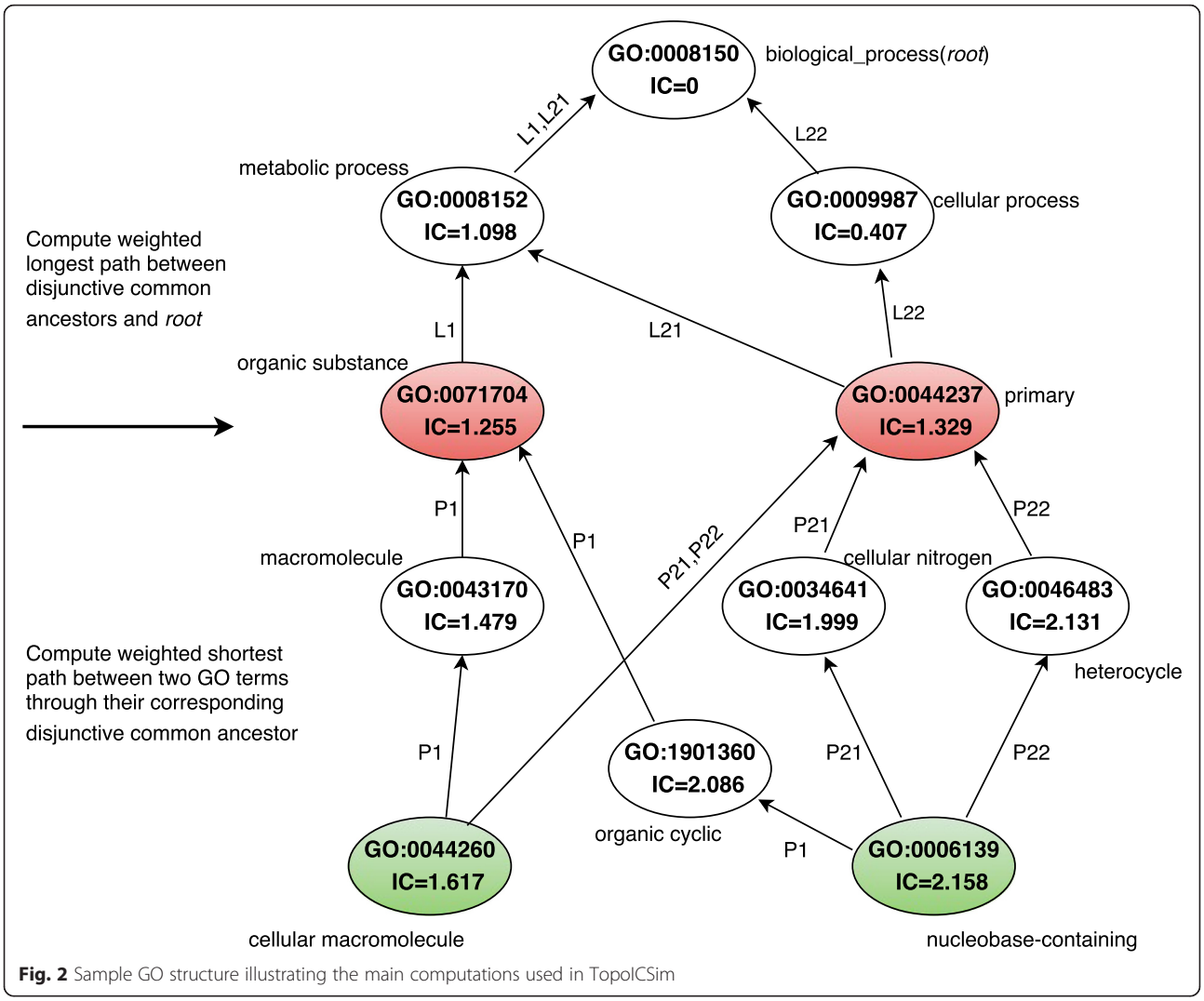

Fig. 2 Sample GO structure illustrating the main computations used in TopoICSim

$$S(\mathrm{GO{:}0044260}, \mathrm{GO{:}0006139}) = 1 - \frac{Arctan(1.076)}{\pi/2} = 0.477$$

Benchmarking of TopoICSim

With the growing number of similarity measures, an important issue is comparison of their performance. For this, in particular the five similarity measures presented in the introduction were considered for comparison with TopoICSim in several tests.

IntraSet similarity

The *IntraSet* similarity is the average similarity over all pairwise comparisons within a set of genes. The *IntraSet* values were calculated with TopoICSim and five other algorithms, namely IntelliGO, Wang, Lord-normalized, Al-Mubaid, and SimGIC, using data sets defined by Pfam clans and KEGG pathways. The performance results obtained with the Pfam clans using MF annotations are shown in Fig. 3. For 7 out of 10 Pfam clans, the TopoICSim measure showed generally higher *IntraSet* similarity compared to the other measures, and only for the CL0289.3 case did it show lower performance. The results for the KEGG pathway datasets based on BP annotations were very similar (Fig. 4). Again the TopoICSim measure had in general higher performance compared to the other measures (11 out of 13).

Discriminating power

The Discriminating Power (*DP*) is defined as the ratio of the *IntraSet* and *InterSet* average gene similarities, where *InterSet* similarities are between gene sets, rather than within. The calculated *DP* values for all methods on the two benchmark datasets used for *IntraSet* similarity are plotted in Figs. 5 and 6. For the Pfam Clans and MF annotations TopoICSim measure was superior compared

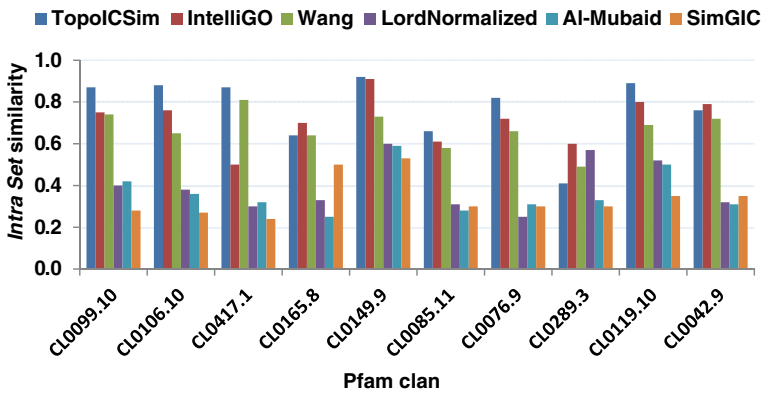

Fig. 3 *IntraSet* similarities for the Pfam clan dataset using MF annotations. The *IntraSet* similarity is estimated for all pairs of genes within in each clan using MF annotations over all considered similarity measures

to the other methods. The minimum and maximum *DP* values generated by the TopoICSim were 1.4 for CL0042.9 and 4.2 for CL0165.8, respectively. For the KEGG pathway dataset the Wang measure provide better performance compared to IntelliGO and TopoICSim, which came second and third.

IntraSet discriminating power

IntraSet Discriminating Power (*IDP*) represents a combination of the *InstraSet* similarity and *DP*, as both should be high for an optimal measure. The *IDP* values were estimated for all measures in the study using formula (16). The results are plotted in Figs. 7 and 8 for MF and BP annotations respectively. For the MF annotations for Pfam clan data TopoICSim shows a generally better performance compared to the other measures. For the BP annotations for KEGG pathway data the best performance was seen for the TopoICSim, IntelliGO, and Wang measures. The TopoICSim had best performance (unique or shared best) for 10 out of 13 cases. It therefore shows a very good and robust performance in this part of the evaluation.

Evaluation versus expression similarity

For evaluation of TopoICSim with respect to annotation similarity associated with expression similarity we used three subsets of human genes from [45], namely G2M, DNA_REPAIR, and STAT3. For each subset both expression and annotation similarities were calculated using Pearson and Spearman correlations and *DC* for expression similarity based on CAGE data (see Methods) (r values), and TopoICSim, IntelliGO, and Wang for semantic similarity (s values). The Self-Organizing Map (SOM) algorithm was used to cluster all interactions into three subsets based on (r, s) values. A 6×6 square topology was selected to set up the SOM computation. The correlation was computed for each cluster and the clusters with $r >= 0.5$ were used to estimate final correlation between expression and annotation similarities as an average on the correlation values within these selected clusters. Table 2 presents the correlation values for each of the three subsets and the considered (r, s) pairs. For the three sets of genes that were tested the maximum correlation was seen when we used the *DC* correlation and TopoICSim measures for the expression and

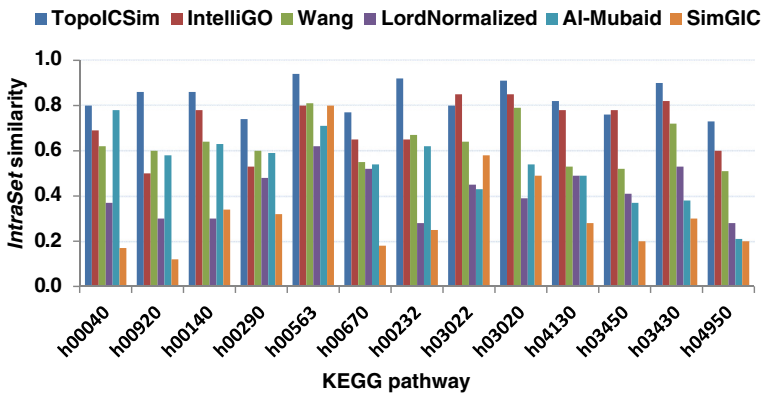

Fig. 4 *IntraSet* similarities for KEGG pathways dataset using BP annotations. The *IntraSet* similarity is estimated for all pair genes within each KEGG pathway using BP annotations for all considered similarity measures

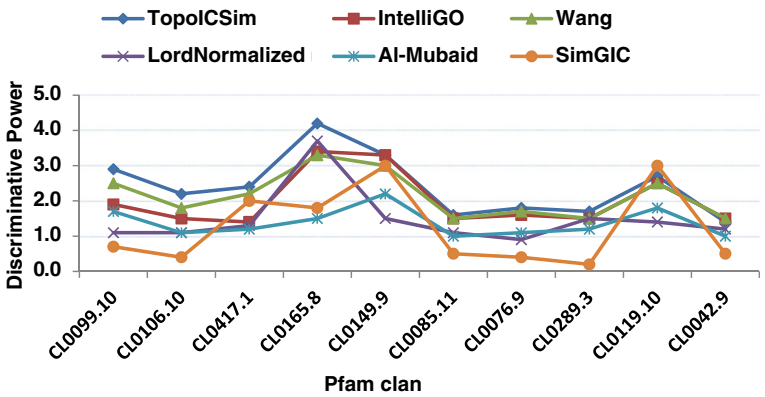

Fig. 5 Comparison of the discriminating power of six similarity measures using Pfam clan and MF annotations. The discriminating power values estimated using all considered similarity measures are plotted for all Pfam clans

annotation similarities (0.943, 0.921, and 0.890 for G2M, DNA_REPAIR, and STAT3 respectively). Also, the calculated correlations with the TopoICSim measure were higher than the correlation values calculated by the two other measures for all cases except the DNA_REPAIR set when using the Spearman and IntelliGO combination (0.89).

Evaluation by CESSM

The TopoICSim measure was used to calculate similarities for the benchmark set of protein pairs downloaded from the CESSM website [34]. The benchmark set represents three different types of similarities, based on sequence similarity (SeqSim), enzyme classification (ECC), and protein domains (Pfam). The results obtained (correlation coefficients) are presented in Table 3. When we used the MF annotations, the correlation coefficients range from 0.55 for the SeqSim dataset to 0.75 for the ECC dataset. The TopoICSim correlation coefficient for the ECC dataset is higher than all other methods. For the Pfam dataset TopoICSim is at a similar level as SimGIC (0.62 vs. 0.63). For the SeqSim dataset the value obtained with TopoICSim is beaten by four other methods (SimGIC, SimUI, RB, LB).

For the BP annotations, the performance was generally higher than for MF annotations. For the ECC and Pfam datasets the TopoICSim correlation coefficients are higher than for any of the other measures. For the SeqSim dataset the score obtained by TopoICSim is beaten by three other measures (SimGIC, SimUI, and RB).

Annotation length bias

Annotations are not uniformly distributed among the genes or gene products within an annotation corpus, and some studies have indicated a clear correlation between semantic scores and the number of annotations [46]. Wang et al. [47] used randomly selected pairs of term groups to evaluate the increase in protein semantic similarity score that resulted only from the increased annotation length, regardless of other biological factors. First, they randomly selected 10,000 pairs of term groups with the same sizes (corresponding to the annotation

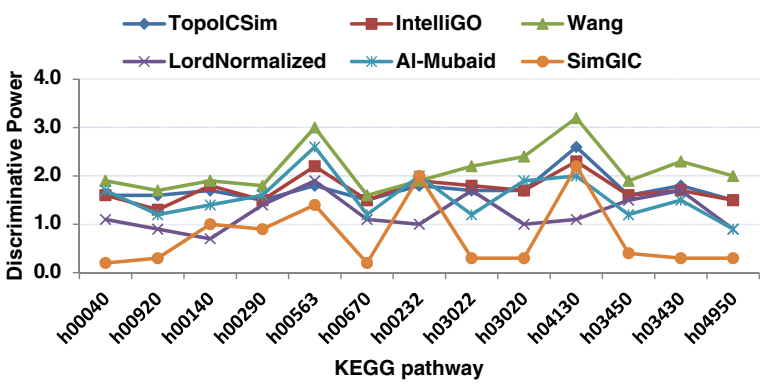

Fig. 6 Comparison of the discriminating power of six similarity measures using KEGG pathway and BP annotations. The discriminating power values estimated with all considered similarity measures are plotted for all KEGG pathways

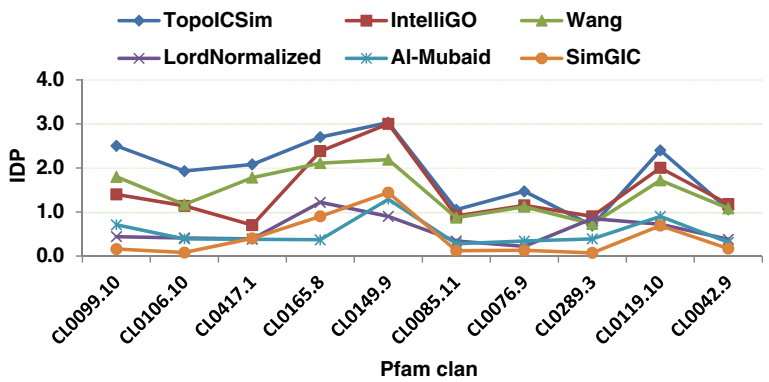

Fig. 7 Comparison of the *IDP* values of six similarity measures using Pfam clan and MF annotations

lengths of proteins) ranging from 1 to 10. Then, using each of 14 semantic similarity scores, they calculated the semantic similarity scores for random term group pairs, and analyzed whether these scores increased as the group size increased using the Spearman rank correlation coefficient. All the 14 semantic similarity methods tested by Wang et al. showed a perfect or close to perfect Spearman correlation (r from 0.99 to 1.00, *p*-value from 9.31e-08 to <2.20e-16). We used their approach and got a Spearman correlation of $r = 0.70$ with *p*-value = 0.02. Although there still is a significant correlation, it is smaller than all reported correlations in Wang et al.

The shallow annotation problem

Genes that are annotated at only very shallow levels (for example "binding") can lead to very high semantic similarities [46]. For example, consider the two human genes Akap1 (A-kinase anchor protein 1 – ID:8165) and Bbs9 (Bardet-Bieddl syndrome 9 – ID:27241). The first gene is a trans-membrane protein that has 10 GO terms associated with the MF ontology. The second gene is poorly understood and has only two GO terms, including GO:0005515 (protein binding), which it happens to share with Akap1. Despite this weak link, some node based methods like Lin and Jiang not only predict high similarity, but actually return a maximum score (1.0). The similarity of these genes according to IntelliGO and Wang is 0.763 and 0.643, respectively, whereas TopoICSim generates a more appropriate low similarity of 0.5.

Running time

Table 4 shows the running times for TopoICSim compared to IntelliGO and Wang, using calculation of the similarity values of all gene pairs in three gene sets that were used for benchmarking. It is not surprising that the Wang method has very short running times compared to TopoICSim and IntelliGO, as Wang does not spend time on finding longest and shortest paths. However, the results also show that TopoICSim actually has shorter running time than IntelliGO in each of the tree cases.

Discussion

Semantic similarity measures rely upon the quality and completeness of their assigned ontology and annotation corpus. The irregular nature of GO annotation data, for example variable edge lengths (edges at the same level can have different semantic measure), variable depth (terms at the same level can have different level of detail), and

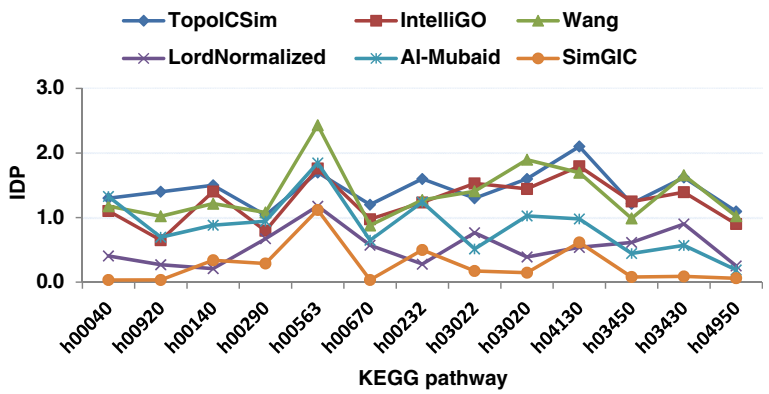

Fig. 8 Comparison of the *IDP* values of six similarity measures using KEGG pathways and BP annotations

Table 2 Correlation between expression and annotation similarities

	G2M			DNA_REPAIR			STAT3		
	TopoICSim	IntelliGO	Wang	TopoICSim	IntelliGO	Wang	TopoICSim	IntelliGO	Wang
Pearson	0.932	0.572	0.849	0.890	0.879	0.867	0.833	0.795	0.824
Spearman	0.914	0.548	0.871	0.876	0.890	0.813	0.872	0.766	0.793
DC	**0.943**	0.594	0.885	**0.921**	0.887	0.863	**0.890**	0.801	0.827

Numbers in bold indicate the best correlation for each subset when comparing TopoICSim, IntelliGO and Wang

variable node density (some areas of the ontology have a larger density of terms than others) should be taken into account by semantic similarity measures.

Most existing methods use in the first step the last (deepest) common ancestor to define similarity between two GO terms, which does not guarantee the shortest path between terms that pass from this common ancestor (i.e. a common ancestor located at a higher level leads to a shorter path between the terms). To overcome this issue TopoICSim measures similarity between two GO terms for all disjunctive common ancestors with the described criteria, and the final similarity measure is returned as the best among them according to (34). Although there are other studies that use disjunctive common ancestors [48], they are node based methods that only use shared information on the disjunctive common ancestors and they do not deal with optimal paths in a subgraph of nodes. Another advantage of the TopoICSim measure is the weighting scheme, which is used according to (29, 30). It leads to a better ability to distinguish between terms with the same semantic similarity but at different levels.

Various strategies have been applied to test the validity of semantic similarity measures [16]. For example, in a gene product interaction network, a functional module is a set of interacting gene products that share a biological process or pathway [46]. Based on this they should display similar MF or BP annotations. This hypothesis was tested by Lord et al. by estimating the correlation between gene annotation (MF annotation) and sequence similarity in set of human proteins [3], since sequence similarity often is associated with functional similarity. Also Guo et al. performed an analysis on all pairs of proteins belonging to the same pathway, which showed higher similarity scores than expected when using BP annotation [49].

For evaluation of the TopoICSim similarity measure in this paper, two benchmarking datasets based on KEGG pathways and Pfam clans were used. These datasets have been obtained directly from [22]. The *IntraSet* similarity, Discriminating Power, and *IntraSet* Discriminating Power values were used for the evaluation. For all quality measures used to evaluate the estimated semantic similarity for these two benchmarking data sets TopoICSim had the best result, except for *DP* values for the KEGG dataset where the Wang method had best performance.

Another common scenario for testing the validity of semantic similarity measures is by testing their correlation with gene expression data. Two gene products with similar function are more likely to have similar expression profile and share same or similar GO terms. Therefore a correlation between gene expressions of two gene products versus the semantic similarity measures can be used as a performance test. Wang et al. [50] compared semantic similarity to expression profile correlation for pairs of genes from the Eisen dataset [51]. They showed that for all the considered measures, high semantic similarity is associated with high expression correlation. Also Sevilla et al. showed correlation between semantic similarity and expression profile, but they dramatically improved it by using grouped data [15]. We took this one step further by applying a SOM algorithm to clustering of gene products by expression

Table 3 Results obtained with the CESSM benchmarking tool

Metrics		Methods											
		SimGIC	SimUI	RA	RM	RB	LA	LM	LB	JA	JM	JB	TopoICSim
MF	ECC	0.62	0.63	0.39	0.45	0.60	0.42	0.45	0.64	0.34	0.36	0.56	**0.75**
	Pfam	**0.63**	0.61	0.44	0.18	0.57	0.44	0.18	0.56	0.33	0.12	0.49	0.62
	SeqSim	**0.71**	0.59	0.50	0.12	0.66	0.46	0.12	0.60	0.29	0.10	0.54	0.55
BP	ECC	0.39	0.40	0.30	0.30	0.44	0.30	0.31	0.43	0.19	0.25	0.37	**0.46**
	Pfam	0.45	0.45	0.32	0.26	0.45	0.28	0.20	0.37	0.17	0.16	0.33	**0.51**
	SeqSim	**0.77**	0.73	0.40	0.30	0.73	0.34	0.25	0.63	0.21	0.23	0.58	0.68

Pearson correlation coefficients are shown for the ECC, Pfam, and SeqSim datasets. The MF and BP annotations are used. Numbers in bold show the best correlation for each dataset. The column headings represent the following methods: *SimGIC* Similarity Graph Information Content, *SimUI* Union Intersection similarity, *RA* Resnick Average, *RM* Resnick Max, *RB* Resnick Best match, *LA* Lord Average, *LM* Lord Max, *LB* Lord Best match, *JA* Jaccard Average, *JM* Jaccard Max, *JB* Jaccard Best match

Table 4 Running time

		Running time (min)		
Gene set	Interactions	TopoICSim	IntelliGO	Wang
STAT3	7569	112	132	15
DNA_REPAIR	22801	312	426	45
G2M	40000	595	815	83

Running times in minutes for calculating similarities over all genes pairs in each of the gene sets

and semantic similarities to select clusters with high correlation. The TopoICSim was superior on the three tested datasets compared to all other similarity measures. Finally, the evaluation with CESSM showed that the TopoICSim is a competitive measure relative to SimGIC, which is superior to all other similarity measures in the CESSM test. However, in the other tests SimGIC had a more variable and sometimes very low performance, which means that TopoICSim in general is a more robust similarity measure with a very good overall performance.

The robust performance was confirmed when we tested for annotation length bias, which has been identified as a potential problem for semantic similarity methods [46]. The analysis showed that although the score still showed some dependency on the number of annotations, the dependency in TopoICSim was clearly lower than for other semantic similarity methods that have been tested. Another potential problem is related to shallow annotation, where high-level GO terms may lead to an overestimation of the similarity between genes. Here TopoICSim should be more robust to such bias than most other methods, due to its design. We have illustrated this with a simple example. Finally, a benchmarking of running time for TopoICSim showed good performance compared to IntelliGO.

Conclusions

In this study we present an improved method for semantic similarity which counts distribution of *IC* on the shortest paths between GO terms and longest path from root to the common ancestors, weighted by their lengths. Several strategies were applied to evaluate the TopoICSim similarity measure. Our results show that the TopoICSim similarity measure is robust, in particular since it was among best similarity measures in all benchmarking tests performed here.

Abbreviations

BP, biological process; CC, MF, cellular component; DC, distance correlation; DP, discriminating power; EC, evidence code; GO, gene ontology; GOA, gene ontology annotation; IC, information content; LCA, lowest common ancestor; MF, molecular function; rDAG, rooted directed acyclic graph; SOM, self-organizing map; SP, shortest path

Acknowledgements

None.

Funding

This work was supported by funding from the Faculty of Medicine, Norwegian University of Science and Technology (NTNU) to RE.

Availability of data and material

All datasets supporting the conclusions of this article are available from open sources and publications as specified in the main text. The TopoICSim script is available for download from http://bigr.medisin.ntnu.no/tools/TopoICSim.R.

Authors' contributions

FD initiated and supervised the project. RE developed, implemented and tested the TopoICSim method and drafted the manuscript. Both authors wrote and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

All data used in this project are from open sources, and do not require ethics approval or consent.

Author details

[1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, P.O. Box 8905, NO-7491 Trondheim, Norway. [2]Department of Mathematics, University of Zabol, Zabol, Iran.

Received: 12 March 2016 Accepted: 21 July 2016
Published online: 29 July 2016

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25(1):25–9.
2. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009–an integrated gene ontology annotation resource. Nucleic Acids Res. 2009;37(Database issue):D396–403.
3. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics. 2003;19(10):1275–83.
4. Ovaska K. Using semantic similarities and csbl. go for analyzing microarray data. Methods Mol Biol. 2015;10:1–12.
5. Meng J, Li R, Luan Y. Classification by integrating plant stress response gene expression data with biological knowledge. Math Biosci. 2015;266:65–72.
6. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Inform. 2012;45(2):363–71.
7. Wu X, Zhu L, Guo J, Zhang DY, Lin K. Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations. Nucleic Acids Res. 2006;34(7):2137–50.
8. Rogers MF, Ben-Hur A. The use of gene ontology evidence codes in preventing classifier assessment bias. Bioinformatics. 2009;25(9):1173–7.
9. Akmal S, Shih L-H, Batres R. Ontology-based similarity for product information retrieval. Computers in Industry. 2014;65(1):91–107.
10. Garla VN, Brandt C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatics. 2012;13:261.
11. Tversky A. Features of similarity. Psychol Rev. 1977;84:327–52.
12. Blanchard E, Harzallah M, Kuntz P. A generic framework for comparing semantic similarities on a subsumption hierarchy, 18th European conference on artificial intelligence (ECAI). 2008. p. 20–4.
13. Wu Z, Palmer M. Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on association for computational linguistics Morristown, NJ, USA: association for computational linguistics. 1994. p. 133–8.
14. Lin D. An information-theoretic definition of similarity. In: ICML '98 proceedings of the fifteenth international conference on machine learning San Francisco, CA, USA: Morgan Kaufmann publishers Inc. 1998. p. 296–304.
15. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ, Rubio A. Correlation between gene expression and GO semantic similarity. IEEE/ACM Trans Comput Biol Bioinform. 2005;2(4):330–8.

16. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009;5(7):e1000443.
17. Shen Y, Zhang S, Wong HS, Zhang L. Characterisation of semantic similarity on gene ontology based on a shortest path approach. Int J Data Min Bioinform. 2014;10(1):33–48.
18. Alvarez MA, Qi X, Yan C. A shortest-path graph kernel for estimating gene product semantic similarity. J Biomed Semantics. 2011;2:3.
19. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Ijcai-95 - proceedings of the fourteenth international joint conference on artificial intelligence, vol. 1 and 2. 1995. p. 448–53.
20. Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the international conference research on computational linguistics. 1997. p. 19–33.
21. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
22. Benabderrahmane S, Smail-Tabbone M, Poch O, Napoli A, Devignes MD. IntelliGO: a new vector-based semantic similarity measure including annotation origin. BMC Bioinformatics. 2010;11:588.
23. Nagar AA-MH. A new path length measure based on go for gene similarity with evaluation using sgd pathways. In: Proceedings of IEEE international symposium on computer-based medical systems. 2008. p. 590–5.
24. Pesquita C, Faria D, Bastos H, Ferreira AE, Falcao AO, Couto FM. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics. 2008;9 Suppl 5:S4.
25. The Sanger Pfam database [http://pfam.xfam.org/]. Accessed 26 July 2016.
26. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
27. The FANTOM5 database [http://fantom.gsc.riken.jp/5/data/]. Accessed 26 July 2016.
28. Song X, Li L, Srimani PK, Yu PS, Wang JZ. Measure the semantic similarity of GO terms using aggregate information content. IEEE/ACM Trans Comput Biol Bioinform. 2014;11(3):468–76.
29. Xu T, Du L, Zhou Y. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. BMC Bioinformatics. 2008;9:472.
30. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern. 1982;43(1):59–69.
31. Székely GRM, Bakirov N. Measuring and testing dependence by correlation of distances. Ann Stat. 2007;35:2769–94.
32. Guo X, Zhang Y, Hu W, Tan H, Wang X. Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. Plos One. 2014;9(2):e87446.
33. de Siqueira SS, Takahashi DY, Nakata A, Fujita A. A comparative study of statistical methods used to identify dependencies between gene expression signals. Brief Bioinform. 2014;15(6):906–18.
34. The Collaborative Evaluation of Semantic Similarity Measures tool [http://xldb.di.fc.ul.pt/tools/cessm/]. Accessed 26 July 2016.
35. Pesquita C, Pessoa D, Faria D, Couto FM. CESSM: Collaborative Evaluation of Semantic Similarity Measures. JB2009: Challenges in Bioinformatics. 2009; 157:190.
36. The ppiPre package [http://cran.r-project.org/web/packages/ppiPre/index.html]. Accessed 26 July 2016.
37. The GOSemSim package [http://bioconductor.org/packages/release/bioc/html/GOSemSim.html]. Accessed 26 July 2016.
38. The SimGIC package [http://csbi.ltdk.helsinki.fi/csbl.go/]. Accessed 26 July 2016.
39. The energy package [http://cran.r-project.org/web/packages/energy/index.html]. Accessed 26 July 2016.
40. The SOMbrero package [http://cran.r-project.org/web/packages/SOMbrero/index.html]. Accessed 26 July 2016.
41. Bioconductor [http://www.bioconductor.org/]. Accessed 26 July 2016.
42. The GOSim package [http://www.bioconductor.org/packages/release/bioc/html/GOSim.html]. Accessed 26 July 2016.
43. The RBGL package [http://www.bioconductor.org/packages/release/bioc/html/RBGL.html]. Accessed 26 July 2016.
44. Sedgewick R, Wayne D. Algorithms. In: Addison-Wesley professional. 2011. p. 661–6.
45. The Hallmark database [http://software.broadinstitute.org/gsea/msigdb/collections.jsp]. Accessed 26 July 2016.
46. Guzzi PH, Mina M, Guerra C, Cannataro M. Semantic similarity analysis of protein data: assessment with biological features and issues. Brief Bioinform. 2012;13(5):569–85.
47. Wang J, Zhou X, Zhu J, Zhou C, Guo Z. Revealing and avoiding bias in semantic similarity scores for protein pairs. BMC Bioinformatics. 2010;11:290.
48. Couto FM, Silva MJ. Disjunctive shared information between ontology concepts: application to gene ontology. J Biomed Semantics. 2011;2:5.
49. Guo X, Liu R, Shriver CD, Hu H, Liebman MN. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics. 2006;22(8):967–73.
50. Wang HAF, Bodenreider O, Dopazo J. Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In: Proceedings of the IEEE symposium on computational intelligence in bioinformatics and computational biology CIBCB 04. 2004. p. 25–31.
51. Eisen MBSP, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95:14863–8.